Identification and initial characterization of a new subgroup in the GH3 gene family in woody plants

Jesús M. Vielba

Department of Plant Physiology, Instituto de Investigaciones Agrobiológicas de Galicia (IIAG-CSIC), Apartado 122, 15780 Santiago de Compostela, Spain. Email: jmvielba@iiag.csic.es

Journal of Plant Biochemistry and Biotechnology
https://doi.org/10.1007/s13562-018-0477-3

Abstract

Plant GH3 proteins are a group of enzymes whose activity is tightly linked with hormone homeostasis and signaling pathways, and therefore their complete characterization is of major relevance for a better understanding of the influence of hormones in plant responses and development programs. According to sequence similarities and substrate specificities, the family is divided in three major subfamilies. In the present work, we have identified a new GH3 Subfamily (IV) with specific sequence characteristics that suggests activity over non-yet described substrates. Our results show that this Subfamily is absent in relevant species like Arabidopsis or maize. The gene promoter analysis of members of this Subfamily in several woody species implies that the activity of these proteins is mainly related to light responses, biotic and abiotic stresses, among other processes. Characterization of the substrates of these enzymes is necessary for a complete understanding of the activity of the GH3 family.

Keywords: GH3 genes · Subfamily IV · Promoter regions · Woody species · *In silico* analysis

Introduction

The Gretchen Hagen 3 (GH3) family of genes, exclusive to the plant kingdom, codes for a group of proteins that directly influences the homeostasis of different plant hormones. GH3 proteins belong to the Firefly luciferase-like acyl-adenylate/thioester forming group of enzymes (Staswick et al. 2002), that is structurally related to the ANL family of adenylating enzymes (Gulick 2009). The GH3 proteins catalyze a two-step reaction where an acyl substrate (hormone) first gets adenylated, and then a nucleophilic attack of the amine group of an amino acid results in the formation of a conjugated molecule of the hormone and the amino acid. Three conserved domains, like those found in the Firefly luciferase family (Chang et al. 1997) are responsible for the catalytic activity of these proteins. Many GH3 proteins show the ability to catalyze reactions using Indole-3-acetic Acid (IAA) and other auxins as substrates. Nevertheless, some GH3 proteins are linked with the activity of Jasmonic Acid (JA), while others can modify benzoates (BA), among which Salicylic Acid (SA) is found (Staswick et al. 2002; Okrent et al. 2009). Nonetheless, the activity of many GH3 proteins has not been characterized yet, as the nature of their substrates remains unknown. In the species where it has been described, the number of GH3 members varies greatly. For example, grapevine has nine proteins, potato has thirteen proteins, Arabidopsis nineteen proteins and soybean twenty-four proteins (Böttcher et al. 2011; Zhang et al. 2018; Okrent and Wildermuth 2011; Yang et al. 2014). Though described in several species, the GH3 family remains to be characterized in many others, particularly in tree species.

Phylogenetic studies have shown that the family is divided into three subfamilies, according to the sequence analysis of the proteins in Arabidopsis and rice (Staswick et al. 2002; Staswick

2005; Terol et al. 2006). This classification is believed to be a resemblance of the substrate specificity of the GH3 proteins (Okrent and Wildermuth 2011). In Subfamily I, AtJAR1 activates JA by catalyzing its conjugation with Ile (Staswick et al. 2002). Several proteins in Subfamily II have shown their ability to catalyze the formation of conjugates between auxins and amino acids, in an apparently redundant mechanism to control the amount of active auxin in the tissues (Staswick 2005; Böttcher et al. 2010). In the Subfamily III, the Arabidopsis AtGH3.12 (PBS3) has been shown to form conjugates between 4-substituted benzoates and amino acids (Okrent et al. 2009). A common feature of these proteins is their ability to accept different amino acids as substrates. In the case of the auxin conjugates, the identity of the amino acid used in the reaction determines whether the resulting molecule is degraded, stored or it becomes a signaling compound (Westfall et al. 2013). Functions of the different proteins are evidently related to the nature of their substrates, but several processes are generally affected by the activity of these proteins, including light and stress responses.

Crystallographic analysis of the GH3 proteins from different subfamilies has shed light on the three-dimensional structure of these proteins. To date, all characterized proteins present a similar folding program, exposing residues in equivalent positions for their interactions with substrates (Westfall et al. 2012; Peat et al. 2012; Round et al. 2013; Westfall et al. 2016). These studies have allowed for a detailed description of the mode of action of these proteins, and to establish the role of the different residues in the two-step reaction that leads to the formation of the conjugates. In the first half-reaction (adenylation), an open conformation of the protein shows an ATP bound, while in the second half-reaction (transfer) a close conformation of the protein shows the binding of an AMP molecule. The definition of the α-helix and the β-strands of the GH3 proteins paved the way for a more specific analysis of the residues involved in the interactions of these proteins with their substrates. According to their biochemical characteristics, eight subgroups were established, with each one of them showing differences in particular residues involved in the interactions with the substrates (Westfall et al. 2012). Despite all these advances in the characterization of the GH3 family, identity of the substrates of several of these enzymes remains unknown.

In the present study, we have described the GH3 family of proteins in several woody species. In this process, we identified and initially characterized a fourth phylogenetic Subfamily, that also establishes a ninth subgroup. Unexpectedly, the Subfamily seems to be absent in many species, including Arabidopsis. Promoter regions from several genes of this newly described group were analyzed to identify putative Transcription Factors Binding Sites (TFBS), in order to infer possible functions of these genes, or the response cascades to which they are related. Only promoter regions from woody species were analyzed here. Our results suggest that these genes might be related to light responses, circadian rhythms, and abiotic and biotic stresses.

Materials and methods

Sequences retrieval

The GH3 protein families from woody species were retrieved from public databases, or from genome sequencing repositories, and manually checked for the presence of false positives, incomplete proteins or the products of pseudogenes. The complete GH3 families from *Citrus clementina*, *Eucalyptus grandis* and *Prunus persica* were retrieved from the Phytozome database (phytozome. jgi.doe.gov), using the keyword "GH3" in the search. *Malus domestica* sequences were obtained from the publication where they were described (Yuan et al. 2013), as it was the case for *Populus trichocarpa* (Okrent and Wildermuth 2011), though in the latter case one missing sequence was added (Böttcher et al. 2011). *Jatropha curcas*, *Populus euphratica*

and *Pyrus bretschneideri* GH3 proteins were obtained from the Pubmed database. *Fraxinus excelsior* proteins were recovered from the webpage of the Ash Tree Genome project (ashgenome.org), *Juglans regia* GH3 family was obtained from the Dendrome webpage (dendrome.ucdavis.edu), and the *Castanea mollissima* GH3 proteins from the Hardwood Genomics database (hardwoodgenomics.org): for these three species the plain text transcriptome database was searched using the sequence of conserved amino acid domains from known GH3 proteins. The proteins were named according to their similitude to the GH3 proteins from *P. trichocarpa* (Okrent and Wildermuth 2011). The reference codes or accession IDs for every protein in the corresponding databases (accessed February 2017) are given in the Supplemental data (ESM-1).

Phylogenetic analysis

The protein sequences were aligned with Muscle (Edgar 2004), and a Maximum Likelihood approach was used for the construction of the tree. This work was developed in the MEGA 7 software (Kumar et al. 2016). A bootstrap analysis was developed by taking 1000 replicates. The supporting values for branches are given in the tree, that was drawn with the Evolview program (He et al. 2016), available in the Evolgenius platform (www.evolgenius. info).

GH3.12 homologs

The CmGH3.12 (*Castanea mollissima* GH3.12) protein was used as the query sequence against the Viridiplantae database to search for homologs in the plant kingdom. The BLASTP (protein–protein Blast) software was used, with default parameters. Only those results with an e value of 0.0 were accepted as homologs, and used thereafter for the characterization of this Subfamily.

Promoter analysis

In order to identify conserved Transcription Factor Binding Sites (TFBS), the 500 base pair (bp) regions upstream of ten GH3.12 genes from different woody species were retrieved from the Pubmed database. When possible, we used the promoter regions from genes of the species we used to build the phylogenetic tree. As this information was not always available, we added promoter regions from identified GH3.12 genes belonging to economically relevant species, like grapevine and cacao tree. We used the 500 bp region upstream from the first codon of the protein, or from the Transcription Start Site (TSS), according to the information available in the NCBI repository. The genes from the following species were used: *Citrus clementina, Eucalyptus grandis, Juglans regia, Malus domestica, Morus alba, Populus euphratica, Populus trichocarpa, Prunus persica, Theobroma cacao* and *Vitis vinifera*. We could not use *J. curcas* because around 200 bp on the promoter region of JcGH3.12 were missing in the Pubmed database.

The MatInspector software (Cartharius et al. 2005) from the Genomatix platform (www.genomatix.de) was used to search for the presence of conserved motifs in the promoter regions, using the PLACE library as database (Higo 1998). Only results showing a probability value of 0.9 or above were accepted. With this value, the presence of mismatches in the sequence of the TFBS identified was limited to one or two, according to the length of the motif. Both motifs in the positive and negative strand were accepted. The biological significance of the identified motifs was confirmed with the PlantPAN database version 2.0 (www. http://plantpan2.itps.ncku.edu.tw/). Both databases, PlantPAN and PLACE, were used to assign specific categories to each TFBS. Information from the publications where the TFBS have been described allowed us to link the motifs with specific responses, tissue specific

expression or particular Transcription Factor (TF) families. In this way, the same motif could be assigned to several categories. For example, the motif MYBGAHV was assigned to three different categories: it is Gibberellin (GA)-responsive, it is bound by a TF from the MYB family and it is found in the promoter region of an alpha-amylase gene (Carbohydrate/energy; Gubler et al. 1995).

Results and discussion

Phylogenetic analysis

We searched the transcriptome of Chinese Chestnut for the complete GH3 family of proteins, in order to obtain information in the genus Castanea, as this is one of the species in which our research is focused. We identified a member of the GH3 family that did not group with any of the previously characterized GH3 groups. To find out the extent of this finding and to expand the current knowledge about GH3 genes in trees, we gathered information of the GH3 family from eleven different woody species. To ease the analysis, the proteins identified as putative GH3 members where named according to their homology with *P. trichocarpa* GH3 proteins (Felten et al. 2009; Okrent and Wildermuth 2011). The only exception was apple, because the GH3 family in this species had been already named and analyzed (Yuan et al. 2013). This nomenclature was established for clarity of the present analysis, and does not intend to be a definitive classification. Those proteins that did not have the three ATP binding domains or did not match the expected size for these proteins were excluded from the analysis. The number of GH3 proteins ranged from eight in *P. persica* to fourteen in *P. euphratica* and *P. trichocarpa*. The proteins from the eleven species where used to build a phylogenetic tree with the Maximum Likelihood method (Fig. 1). The GH3 proteins formed five clusters in the tree, with some of them representing already described subfamilies (I, II and III), while two of them appeared as new ones (II–III and IV). All species showed to have at least one protein in every cluster except for *P. bretschneideri* and *F. excelsior* in Subfamily IV. As previously stated, Subfamily I is divided in two subgroups (Wakuta et al. 2011), and every species showed to have at least one member in each of these subgroups (Fig. 1).

The phylogenetic identity of Subfamily IV proteins was clearly established in our analysis. We decided to name these proteins GH3.12 because this codename was absent in *P. trichocarpa*. The GH3.12 proteins from distantly-related woody species grouped together in the tree, while they were clearly separated from other GH3 proteins from the same species (Fig. 1). One member of the newly described Subfamily IV from this tree had previously been identified in other species. For *M. domestica*, authors found that MdGH3.15 was significantly different from the rest of GH3 members of the family, and it showed to be the only member without an homolog in the apple genome (Yuan et al. 2013). Okrent and Wildermuth (2011) failed to identify PtGH3.12 in their analysis because they used Arabidopsis genes as seeds in their search, and this Subfamily is not present in this model species. Nonetheless, this protein had previously been identified though no further characterized (Böttcher et al. 2011). We have found that all of these proteins belong to a new Subfamily that is widely spread in the plant kingdom, but that is absent in several relevant species.

All species showed at least one protein in subfamilies II and III, and in the mixed Subfamily II–III. Apparently, the method we used to build the tree was unable to clearly split the previously established Subfamilies II and III, leaving in a separate group a set of proteins that had been seen to match in Subfamily II (Okrent and Wildermuth 2011; Yuan et al. 2013). Indeed, Subfamily II–III showed to be more related to Subfamily III than to Subfamily II (Fig. 1), suggesting a more complicated phylogenetic relationship among GH3 Subfamilies than

previously stated. Furthermore, apple had been shown to lack proteins in Subfamily III (Yuan et al. 2013), while our analysis placed four GH3 proteins in that group. The Arabidopsis protein AtGH3.5, which is an homolog of the proteins in the II-III group, has been shown to be active on hormones from different families, including substrates from groups II and III (auxins and SA; Westfall et al. 2016). Therefore, there exists the possibility that indeed proteins in this group might constitute a fifth Subfamily of GH3 proteins, though research on other members of this putative group is necessary. So far, we decided to leave these proteins as a mixed group between Subfamilies II and III.

GH3.12 homologous proteins

We blasted the CmGH3.12 sequence against the NCBI repository to find out the range of plant species in which GH3.12 homologs could be found (ESM-2). The results showed that homologous proteins existed in several species, including woody and non-woody species, as well as monocot and eudicot species. In total, we identified 45 species with close homologs to CmGH3.12 (e=0.0), with some genus showing two or three species (i.e. Oryza, Gossypium). As suspected, Arabidopsis did not show to have homologs in this Subfamily. Some of the proteins identified with Blast had been previously included in other studies. In one of the first efforts to characterize GH3 families in model species, Terol et al. (2006) found that OsGH3.6 did not match within any of the three established subfamilies, but did not further analyze this issue. A recent study of the *Medicago truncatula* GH3 family showed the presence of three proteins from this Subfamily in this species (only two functional), leading the authors to suggest that it might be a species-specific subgroup (Yang et al. 2014). Accordingly, Böttcher et al. (2011) also identified proteins from this Subfamily in *P. trichocarpa* and *V. vinifera*, but they did not characterize them. The search revealed that many other species lack genes in this Subfamily, including several whose genomes have been completely sequenced. For example, no genes from Subfamily IV could be found in maize, potato, tobacco or the family Brassicaceae.

Biochemical in silico characterization

Three identified GH3.12 proteins (CmGH3.12, PtGH3.12 and VviGH3.12) were aligned with GH3 proteins from the eight different subgroups previously defined according to their biochemical characteristics (Westfall et al. 2012). We included one member from each subgroup, in order to analyze differences in conserved domains and relevant protein regions (Fig. 2). With respect to the three main conserved domains, that interact with the nucleotide in the conjugation reaction, Domain 3 did not show particular differences with other subfamilies. On the other hand, Domain 1 showed the presence of significant differences in the second half of the domain, though essential residues were conserved. Domain 2 of proteins from Subfamily IV presented a Phe residue in the second position (^{339}Y<u>F</u>ASE^{343}, numeration corresponds to CmGH3.12) that had not been characterized in any other GH3 protein. This Phe residue is a Leu residue in rice species (see gene references in ESM-2). At this position, other GH3 proteins showed amino acids with small side-chains, like Ala, Gly or in rare cases Val (Westfall et al. 2012). This finding might be relevant because of the interactions in which this domain is involved, with both ATP/AMP, and with the specific substrate of the conjugation reaction (Peat et al. 2012; Singh et al. 2015). The presence of the aromatic ring of Phe might be influencing the outcome of the reaction because of the steric hindrance generated. By means of mutational substitution, Westfall et al. (2016) changed this residue in the sequence of AtGH3.5, from Ala (R group— CH3) to Val (R group—3HC–CH–CH3). This substitution rendered the protein mostly inactive on auxin, one of the known substrates of AtGH3.5, what the authors suggested was a consequence of steric hindrance generated by the greater R group from Val (Westfall et al. 2016). Therefore, we suggest that the voluminous R group from Phe might also be an

impediment for the interaction with auxins carrying an indole ring, and the substrates should have to be different and, probably, smaller molecules.

Many amino acids that are involved in the interaction with the hormone substrate were different in GH3.12 proteins with respect to the rest of biochemically characterized subgroups. The GH3.12 residues in α-helix 5 were strikingly different from those in other subgroups, and almost conserved within the Subfamily IV proteins (Fig. 2b). Residues surrounding Domain 2 (^{337}GD^{338}, ^{344}CC^{345}) were notoriously different in GH3.12 with respect to the other proteins in the alignment. For α-helix 6 some similarities could be found, mainly with subgroups 1, 3, 4 and 8. Nonetheless, other residues involved in acyl binding, like Gly^{331} and Trp^{336} in AtGH3.11, were not conserved in GH3.12 proteins, suggesting different acyl substrates for these proteins (Fig. 2; Peat et al. 2012). Therefore, the GH3.12 proteins seem to retain the ability to bind ATP/ AMP and catalyze the formation of conjugates, but the substrates described for other GH3 proteins might not be able to fit within the active site of the GH3.12 proteins.

Nonetheless, the GH3 proteins have shown a significant promiscuity in their ability to bind related yet different substrates. For instance, Arabidopsis proteins from Subfamily II have shown to be active *in vitro* on several auxin substrates (Staswick 2005), as have also been seen *in silico* for other GH3 proteins (Vielba et al. 2016).

Promoter analysis

To gain further insight into the putative functions of these proteins, we developed an analysis of the promoter regions of several of the *GH3.12* genes from woody species, in an effort to identify conserved Transcription Factors Binding Sites (TFBS). The number of TFBS found ranged from thirty-seven in *E. grandis* to sixty-three in *V. vinifera*. The complete list of identified motifs can be seen in ESM-3, and relevant results are presented altogether in Table 1. We used 500 bp from the promoter region of the corresponding genes of those GH3.12 proteins previously used in the phylogenetic tree, when that information was available. When it was not possible, we used the promoter region of other identified GH3.12 proteins from woody species, like VviGH3.12 or PpeGH3.12. Both motifs found in the plus and the minus strand were taken into consideration, as they are believed to be equally important (Lis and Walther 2016). The motifs found were classified by the following criteria: first, they were classified according to the kind of stimulus to which they respond or the organ-specific expression conferred by them; second, by the presence of hormone-specific responsive motifs, and last by the family of the TFs that putatively bind them. Overall results suggest that GH3.12 proteins are indeed implicated in same functions as previously characterized GH3 proteins, i.e. light responses, biotic and abiotic stresses, and growth (Wang et al. 2008; Okrent and Wildermuth 2011). Though every promoter region was different, some similarities could be found, like the preponderance of motifs related to light and phytochrome responses, circadian rhythms and stresses from different nature. Circadian rhythms might be the main function not previously assigned to the GH3 family that was highlighted from our results; nonetheless, it is not rare, as it is believed that around one third of the plant genes are under the control of circadian rhythms (Covington et al. 2008). Besides, tissue-specific expression was found to be common for all regions analyzed, in particular for root tissues and seed or embryo structures. Moreover, metabolism (where we included energy and carbohydrate-related motifs) showed to be relevant for all species. Whether the identified motifs exert a positive or negative effect on gene expression remains to be investigated. It is important to keep in mind the combinatorial nature of TFs (Brkljacic and Grotewold 2017), as several of them will interact to effectively activate (or repress) the machinery for the transcription of the gene.

Environmental cues

Light and phytochrome related TFBS (grouped together) showed to be the most relevant stimulus controlling the expression of the genes in this Subfamily belonging to woody species, with as many as eighteen motifs detected in *P. euphratica*, seventeen in *M. domestica* and fifteen in *P. trichocarpa*, *V. vinifera* and *J. regia*. *P. persica* showed the lowest number with seven motifs. Also, circadian rhythms-related TFBS were found in all promoter regions. Particularly, the motif "Ciacadianlelhc" (CAANNNNATC; Piechulla et al. 1998) was found in all the regions analyzed, from two copies in *P. persica* and *C. clementina* to twelve copies in *M. domestica*. Biotic and abiotic stresses showed to have a strong influence on the expression of these genes. For example, the sulfur-deficiency related motif "Surecoreatsultr11" (Maruyama-Nakashita et al. 2005) was found in six of the promoters. The "Oserootnodule" motifs (1 and 2), related to expression in infected cells of root nodules (Vieweg et al. 2005) were found in seven of the promoter regions analyzed. Furthermore, a motif related to both biotic and abiotic stress because of its responsiveness to salt and pathogens ("Gt1gmscam4"; Park et al. 2004) was also detected in seven of the promoter regions.

Hormone-related TFBS

The major differences between the promoter regions analyzed concerned the identification of TFBS related to hormones. Only binding sites responsive to GAs could be found in every promoter region, but in a low number for most species. The ABA and AUX responsive motifs were found in eight out of ten regions analyzed, but the number varied among species. A particular case was *T. cacao*, that did not show responsive elements to ABA or AUX. In seven of the ten regions analyzed we found motifs related to ET and SA, but in a low number except for ET in both Populus species and *C. clementina*, and SA in *V. vinifera*. The overall number of hormone-responsive motifs found in these promoter regions ranged from seven in *T. cacao* to seventeen in *P. trichocarpa*.

The TFBS can have a positive or negative effect on gene expression, depending on several factors. In an analysis of the promoter region of *GH3* genes in tomato, it was found that GA repressed the expression of 12 of the 15 *GH3* genes in this species (Kumar et al. 2012). We identified GA-related motifs in every promoter. Therefore, more research is needed to clarify if the effect of those motifs is positive or negative on gene expression remains to be investigated. ABA is a stress-related hormone, and its influence on the expression of *GH3* genes has already been appointed (Kumar et al. 2012; Yang et al. 2014).

Auxin motifs were not present in all regions, once again suggesting that these hormones are not the substrates for the GH3.12 proteins. In *GH3* genes from Subfamily II, which are active on auxin substrates, the hormone directly influences the expression of many of these genes (Okrent and Wildermuth 2011; Kumar et al. 2012; Yang et al. 2014). We found less auxin-responsive elements than expected, suggesting that this hormone is not a master regulator of the expression of these *GH3.12* genes. In the analysis of the promoter regions of *GH3* genes in maize, authors found several auxin-related motifs, including ASF1MOTIFCAMV, which the authors identified in all promoter regions but one (Zhang et al. 2016). We identified this motif in some species (one time in *M. domestica* and *M. alba*, two times in *V. vinifera*; ESM-3), but its presence was not extended to all promoter regions of *GH3.12* genes. Our results suggest that auxin is not a significant factor controlling *GH3.12* expression, with the exception of *C. clementina*, where we identified eight auxin-responsive motifs. Together with the suspected inability of indole rings to fit within the active site of GH3.12 proteins, our results suggest that these proteins are not active on auxins.

Transcription factor families

MYB showed to be the most relevant TF family influencing the expression of *GH3.12* genes in these woody species, as binding motifs for these proteins could be found in every promoter analyzed. The MYB genes have been shown to be involved in development and metabolism, as well as in responses to both biotic and abiotic stresses (Dubos et al. 2010). In trees, MYB genes have been implicated in responses to abiotic stress, like salt stress in *P. trichocarpa* or drought stress in *Pyrus betulaefolia* (Fang et al. 2017; Li et al. 2017). In chestnut, a recent analysis concerning resistance of different clones to *Phytophthora cinnamomi* showed that the expression of a MYB gene was directly linked to a higher resistance to the oomycete (Santos et al. 2017). The activity of MYB TFs on GH3.12 promoters is in agreement with a role for these genes in stress responses in these woody species, as they would be part of the signaling cascades activated by MYB proteins. The other family of TFs with presence in all promoters was WRKY, though in a lower number than MYB. The WRKY TFs have been related to (a)biotic stress and seed germination (Tripathi et al. 2014). In the analysis of the promoter regions of *GH3* genes in Arabidopsis, it was found that WRKY, bZIP and MYB transcription factors were among the most relevant TFs binding those regions (Okrent and Wildermuth 2011), in a similar trend with our results. Nonetheless, authors also found a significant presence of TFs from the MYC2 family, which were not identified in our analysis.

Other relevant TF family found in the analysis was the Basic leucine-zipper (bZIP) family. These TFs are related to photomorphogenesis, energy homeostasis, root development, abiotic and biotic stresses (Guedes Correa et al. 2008; Noman et al. 2017). A member of this family in tobacco was shown to directly bind the promoter region of a *GH3* gene, in what the authors suggested was a interaction with other different TFs (Heinekamp et al. 2004). We found an important number of binding motifs for the most relevant TFs in GH3.12 promoters (from 12 to 26 in 500 bp), suggesting that indeed several different factors influence their expression. The relation of *GH3.12* genes with energy and growth control might be a consequence of the control that bZIP genes have over their activity. In Arabidopsis, it has been shown that bZIP-related TFs interfere auxin signaling in the roots to adapt primary root growth to energy availability (Weiste et al. 2017). Activity of GH3.12 proteins over non-identified substrates might be directly influencing whether the plant invests energy in growth or in stress responses, a situation that has been proposed before for other *GH3* genes (Park et al. 2007).

By means of the information present in public repositories, we have identified the GH3 protein families in several woody species, leading to the description of a new Subfamily. It is not exclusive to woody species, and it is absent in several relevant plant genera. Besides, phylogenetic analysis of the complete GH3 protein families from several woody species suggests that actual classification of the family might not be entirely accurate. The analysis of the residues of these enzymes involved in substrate binding suggested that, though the identity of those substrates remains unknown, it might be different from those described for other GH3 proteins. This would exclude auxins, jasmonic acid or benzoates as the acyl substrates used by these adenylating enzymes. The analysis of the promoter region of ten *GH3.12* genes from different woody species provided an idea of the processes an conditions influencing its expression, including circadian rhythms, light, and biotic and abiotic stresses. Likewise, several tissue-specific motifs were also identified, as well as hormone-linked motifs. Further biochemical, syntenic and genetic analysis are needed to fully characterize the Subfamily IV, and by extension the complete GH3 family.

Acknowledgements: The authors would like to thank Dr. Elena Varas for critical reading of the early version of this manuscript. The work is dedicated in loving memory to Brais Bogo, who passed away much too soon.

References

Böttcher C, Keyzers RA, Boss PK, Davies C (2010) Sequestration of auxin by the indole-3-acetic acid-amido synthetase GH3-1 in grape berry (*Vitis vinifera* L.) and the proposed role of auxin conjugation during ripening. *J Exp Bot* 61:3615–3625. https:// doi.org/10.1093/jxb/erq174

Böttcher C, Boss PK, Davies C (2011) Acyl substrate preferences of an IAA-amido synthetase account for variations in grape (*Vitis vinifera* L.) berry ripening caused by different auxinic com-pounds indicating the importance of auxin conjugation in plant development. *J Exp Bot* 62:4267–4280. https://doi.org/10.1093/ jxb/err134

Brkljacic J, Grotewold E (2017) Combinatorial control of plant gene expression. *Biochim Biophys Acta Gene Regul Mech* 1860:31–40 https://doi.org/10.1016/j.bbagrm.2016.07.005

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21:2933–2942. https://doi.org/10.1093/ bioinformatics/bti473

Chang KH, Xiang H, Dunaway-Mariano D (1997) Acyl-adenylate motif of the acyl-adenylate/thioester-forming enzyme superfamily: a site-directed mutagenesis study with the Pseudomonas sp. Strain CBS3 4-chlorobenzoate:coenzyme A ligase. *Biochemistry* 36:15650–15659. https://doi.org/10.1021/bi971262p

Covington MF, Maloof JN, Straume M, Kay SA, Harmer SL (2008) Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol* 9: R130. https://doi.org/10.1186/gb-2008-9-8-r130

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L (2010) MYB transcription factors in Arabidopsis. *Trends Plant Sci* 15(10):573–581. https://doi.org/10.1016/j.tplants.2010.06. 005

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Fang Q, Jiang T, Xu L, Liu H, Mao H, Wang X, Jiao B, Duan Y, Wang Q, Dong Q, Yang L, Tian G, Zhang C, Zhou Y, Liu X, Wang H, Fan D, Wang B, Luo K (2017) A salt-stress-regulator from the Poplar R2R3 MYB family integrates the regulation of lateral root emergence and ABA signaling to mediate salt stress tolerance in Arabidopsis. *Plant Physiol Biochem* 114:100–110. https://doi.org/10.1016/j.plaphy.2017.02.018

Felten J, Kohler A, Morin E, Bhalerao RP, Palme K, Martin F, Ditengou FA, Legue V (2009) The ectomycorrhizal fungus *Laccaria bicolor* stimulates lateral root formation in Poplar and Arabidopsis through auxin transport and signaling. *Plant Physiol* 151:1991–2005. https://doi.org/10.1104/pp.109.147231

Gubler F, Kalla R, Roberts JK, Jacobsen JV (1995) Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter. *Plant Cell* 7(11):1879–1891. https://doi.org/10. 1105/tpc.7.11.1879

Guedes Correa LG, Riaño-Pachón DM, Guerra Schrago C, Vicentini dos Santos R, Mueller-Roeber B, Vincentz M (2008) The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. *PLoS One* 3(8):e2944. https://doi.org/10.1371/journal.pone.0002944

Gulick AM (2009) Conformational dynamics in the acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem Biol* 4(10):811–827 https://doi.org/10.1021/cb900156h

He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res* 44:W236–W241. https://doi.org/10.1093/nar/gkw370

Heinekamp T, Strathmann A, Kuhlmann M, Froissard M, Müller A, Perrot-Rechenmann C, Dröge-Laser W (2004) The tobacco bZIP transcription factor BZI-1 binds the GH3 promoter in vivo and modulates auxin-induced transcription. *Plant J* 38:298–309. https://doi.org/10.1111/j.1365-313X.2004.02043.x

Higo K (1998) PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res* 26:358–359. https://doi.org/10. 1093/nar/26.1.358

Kumar R, Agarwal P, Tyagi AK, Sharma AK (2012) Genome-wide investigation and expression analysis suggest diverse roles of auxin-responsive GH3 genes during development and response to different stimuli in tomato (*Solanum lycopersicum*). *Mol Genet Genom* 287:221–235. https://doi.org/10.1007/s00438-011-0672-6

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874. https://doi.org/10.1093/molbev/ msw054

Li K, Xing C, Yao Z, Huang X (2017) PbrMYB21, a novel MYB protein of *Pyrus betulaefolia*, functions in drought tolerance and modulates polyamine levels by regulating arginine decarboxy-lase gene. *Plant Biotechnol J* 15:1186–1203. https://doi.org/10. 1111/pbi.12708

Lis M, Walther D (2016) The orientation of transcription factor binding site motifs in gene promoter regions: does it matter? *BMC Genom* 17:185. https://doi.org/10.1186/s12864-016-2549-x

Maruyama-Nakashita A, Nakamura Y, Watanabe-Takahashi A, Inoue E, Yamaya T, Takahashi H (2005) Identification of a novel cis-acting element conferring sulfur deficiency response in Arabidopsis roots. *Plant J* 42:305–314. https://doi.org/10.1111/j. 1365-313X.2005.02363.x

Noman A, Liu Z, Aqeel M, Zainab M, Khan MI, Hussain A, Ashraf MF, Li X, Weng Y, He S (2017) Basic leucine zipper domain transcription factors: the vanguards in plant immunity. *Biotechnol Lett* 39(12):1779–11791. https://doi.org/10.1007/s10529-017-2431-1

Okrent RA, Wildermuth MC (2011) Evolutionary history of the GH3 family of acyl adenylases in rosids. *Plant Mol Biol* 76:489–505. https://doi.org/10.1007/s11103-011-9776-y

Okrent RA, Brooks MD, Wildermuth MC (2009) Arabidopsis GH3.12 (PBS3) conjugates amino acids to 4-substituted benzoates and is inhibited by salicylate. *J Biol Chem* 284:9742– 9754. https://doi.org/10.1074/jbc.M806662200

Park HC, Kim ML, Kang YH, Jeon JM, Yoo JH, Kim MC, Park CY, Jeong JC, Moon BC, Lee JH, Yoon HW, Lee SH, Chung WS, Lim CO, Lee SY, Hong JC, Cho MJ (2004) Pathogen- and NaCl-induced expression of the SCaM-4 promoter is mediated in part by a GT-1 box that interacts with a GT-1-like transcription factor. *Plant Physiol* 135:2150–2161. https://doi.org/10.1104/pp. 104.041442

Park JE, Park JY, Kim YS, Staswick PE, Jeon J, Yun J, Kim SY, Kim J, Lee YH, Park CM (2007) GH3-mediated auxin homeostasis links growth regulation with stress adaptation response in Arabidopsis. *J Biol Chem* 282:10036–10046. https://doi.org/10. 1074/jbc.M610524200

Peat TS, Bo¨ttcher C, Newman J, Lucent D, Cowieson N, Davies C (2012) Crystal structure of an indole-3-acetic acid amido synthetase from grapevine involved in auxin homeostasis. *Plant Cell* 24:4525–4538. https://doi.org/10.1105/tpc.112.102921

Piechulla B, Merforth N, Rudolph B (1998) Identification of tomato Lhc promoter regions necessary for circadian expression. *Plant Mol Biol* 38:655–662. https://doi.org/10.1023/A:1006094015513

Round A, Brown E, Marcellin R, Kapp U, Westfall CS, Jez JM, Zubieta C (2013) Determination of the GH3.12 protein conformation through HPLC-integrated SAXS measurements combined with X-ray crystallography. *Acta Crystallogr Sect D Biol Crystallogr* 69:2072–2080. https://doi.org/10.1107/S0907444913019276

Santos C, Duarte S, Tedesco S, Fevereiro P, Costa RL (2017) Expression profiling of Castanea genes during resistant and susceptible interactions with the oomycete pathogen *Phytophthora cinnamomi* reveal possible mechanisms of immunity. *Front Plant Sci*. https://doi.org/10.3389/fpls.2017.00515

Singh VK, Jain M, Garg R (2015) Genome-wide analysis and expression profiling suggest diverse roles of GH3 genes during development and abiotic stress responses in legumes. *Front Plant Sci.* https://doi.org/10.3389/fpls.2014.00789

Staswick PE (2005) Characterization of an Arabidopsis enzyme family that conjugates amino acids to Indole-3-Acetic Acid. *Plant Cell Online* 17:616–627. https://doi.org/10.1105/tpc.104. 026690

Staswick PE, Tiryaki I, Rowe ML (2002) Jasmonate response locus JAR1 and several related Arabidopsis genes encode enzymes of the Firefly luciferase superfamily that show activity on Jasmonic, Salicylic, and Indole-3-Acetic Acids in an assay for adenylation. *Plant Cell* 14:1405–1415. https://doi.org/10.1105/tpc.000885

Terol J, Domingo C, Talón M (2006) The GH3 family in plants: genome wide analysis in rice and evolutionary history based on EST analysis. *Gene* 371:279–290. https://doi.org/10.1016/j.gene.2005.12.014

Tripathi P, Rabara RC, Rushton PJ (2014) A systems biology perspective on the role of WRKY transcription factors in drought responses in plants. *Planta* 239(2):255–266. https://doi.org/10.1007/s00425-013-1985-y

Vielba JM, Varas E, Rico S, Covelo P, Sánchez C (2016) Auxin-mediated expression of a GH3 gene in relation to ontogenic state in Chestnut. *Trees Struct Funct* 30:2237–2252. https://doi.org/10.1007/s00468-016-1449-7

Vieweg MF, Hohnjec N, Küster H (2005) Two genes encoding different truncated hemoglobins are regulated during root nodule and arbuscular mycorrhiza symbioses of *Medicago truncatula*. *Planta* 220:757–766. https://doi.org/10.1007/s00425-004-1397-0

Wakuta S, Suzuki E, Saburi W, Matsuura H, Nabeta K, Imai R, Matsui H (2011) OsJAR1 and OsJAR2 are jasmonyl-L-isoleucine synthases involved in wound- and pathogen-induced jasmonic acid signalling. *Biochem Biophys Res Commun* 409(4):634–639. https://doi.org/10.1016/j.bbrc.2011.05.055

Wang H, Tian CE, Duan J, Wu K (2008) Research progresses on GH3s, one family of primary auxin-responsive genes. *Plant Growth Regul* 56(3):225–232. https://doi.org/10.1007/s10725-008-9313-4

Weiste C, Pedrotti L, Selvanayagam J, Muralidhara P, Fröschel C, Novak O, Ljung K, Hanson J, Dröge-Laser W (2017) The Arabidopsis bZIP11 transcription factor links low-energy signalling to auxin-mediated control of primary root growth. *PLoS Genet* 13(2):e1006607. https://doi.org/10.1371/journal.pgen. 1006607

Westfall CS, Zubieta C, Herrmann J, Kapp U, Nanao MH, Jez JM (2012) Structural basis for prereceptor modulation of plant hormones by GH3 proteins. *Science* 6089:1708–1711. https://doi.org/10.1126/science.1221863

Westfall CS, Muehler AM, Jez JM (2013) Enzyme action in the regulation of plant hormone responses. *J Biol Chem* 288:19304– 19311. https://doi.org/10.1074/jbc.R113.475160

Westfall C, Sherp AM, Zubieta C, Alvarez S, Schraft E, Marcellin R, Ramirez L, Jez JM (2016) *Arabidopsis thaliana* GH3.5 acyl acid amido synthetase mediates metabolic crosstalk in auxin and salicylic acid homeostasis. *Proc Natl Acad Sci* 113:13917–13922. https://doi.org/10.1073/pnas.1612635113

Yang Y, Yue R, Sun T, Zhang L, Chen W, Zeng H, Wang H, Shen C (2014) Genome-wide identification, expression analysis of GH3 family genes in *Medicago truncatula* under stress-related hormones and *Sinorhizobium meliloti* infection. *Appl Microbiol Biotechnol* 99:841–854. https://doi.org/10.1007/s00253-014-6311-5

Yuan H, Zhao K, Lei H, Shen X, Liu Y, Liao X, Li T (2013) Genome-wide analysis of the GH3 family in apple (*Malus domestica*). *BMC Genom* 14:297. https://doi.org/10.1186/1471-2164-14-297

Zhang DF, Zhang N, Zhong T, Wang C, Xu ML, Ye JR (2016) Identification and characterization of the GH3 gene family in maize. *J Integr Agric* 15:249–261. https://doi.org/10.1016/S2095-3119(15)61076-0

Zhang C, Zhang L, Wang D, Ma H, Liu B, Shi Z, Ma X, Chen Y, Chen Q (2018) Evolutionary history of the Glycoside Hydrolase 3 (GH3) family based on the sequenced genomes of 48 plants and identification of Jasmonic Acid-Related GH3 proteins in *Solanum tuberosum*. *Int J Mol Sci* 19(7):1850. https://doi.org/10. 3390/ijms19071850

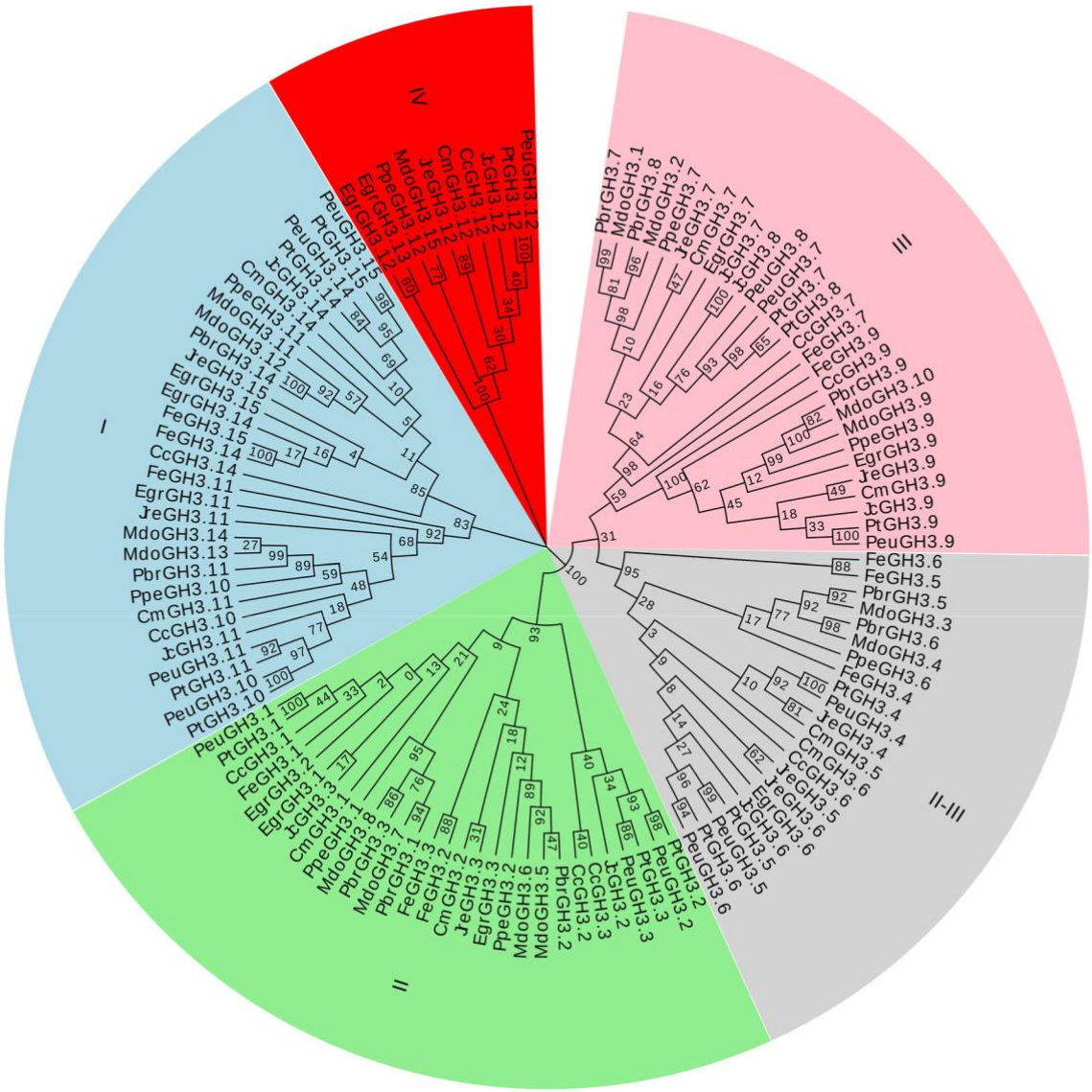

Fig. 1. GH3 phylogenetic tree. Phylogenetic analysis of the GH3 proteins from eleven woody species. The tree was built with the Maximum Likelihood Method. The analysis established five different clusters of proteins (I, II, II–III, III and IV). Bootstrap values are given for every branch. Abbreviations: Cc-*Citrus clementina*; Cm- *Castanea mollissima*; Egr:-*Eucalyptus grandis*; Fe-*Fraxinus excelsior*; Jc-*Jatropha curcas*; Jre-*Juglans regia*; Mdo-*Malus domestica*; Pbr-*Pyrus bretschneideri*; Peu-*Populus euphratica*; Pt-*Populus trichocarpa*; Ppe-*Prunus persica.*

A

		DOMAIN 1	DOMAIN 2	DOMAIN 3
AtGH3.11/JAR1	Group 1	SGTSQGRPK	YGSSE	GLYRYRLGD
AtGH3.3	Group 2	SGTSAGERK	YASSE	GLNRYRVGD
AtGH3.10/DFL2	Group 3	SGTTEGRQK	YGSTE	GLYRYRLGD
AtGH3.12/PBS3	Group 4	SGTSGGAQK	YGSSE	GLYRMRVGD
AtGH3.8	Group 5	SGTSGGKQK	YVSSE	GLHRYKMGD
AtGH3.13	Group 6	SGTTAGIQN	YGSSE	GLYRYRVGD
AtGH3.19	Group 7	TGTSGGIHK	YGSSE	GLHRYLIGD
OsGH3.7	Group 8	SGTSGGQQK	YASTE	GLYRYRVGD
CmGH3.12	Group 9	SGASTMKPK	YFASE	GFYRYRLGD
VvGH3.12	Group 9	SGTSSMKPK	YFASE	GFFRYRLGD
PtGH3.12	Group 9	SGTSTMKPK	YFASE	GLYRYRLGD

B

		α5	α6	230	312	β8-β9
AtGH3.11/JAR1	Group 1	FTDELMENTLQLFRTA	ATTNV	V	I	HDYGSSEGW
AtGH3.3	Group 2	TIDEDMDR-RQLLYSL	VLTSY	V	I	TMYASSESY
AtGH3.10/DFL2	Group 3	FTRHSAQTTLQIFRLS	ATTHY	A	I	ADYGSTESW
AtGH3.12/PBS3	Group 4	WNNKYLDN-LTFIYDL	ATSSY	I	V	TTYGSSETT
AtGH3.8	Group 5	RNNKYLEN-IKFIFYY	AASTS	A	I	SSYVSSETM
AtGH3.13	Group 6	LTTEDGEQRIMFGSLY	MITCI	P	V	SWYGSSECF
AtGH3.19	Group 7	VNDKYIEN-LGYLLAV	SFTSY	T	I	MVYGSSESI
OsGH3.7	Group 8	STAEELDR-KVFFYAV	ALITY	S	I	PIYASTECA
CmGH3.12	Group 9	YFDSKLSKAASFIAHQ	ASSYP	P	V	GDYFASECC
VvGH3.12	Group 9	YFDSPPSKAASHIAHQ	ASAFP	A	V	GDYFASECC
PtGH3.12	Group 9	YFDSALSKAASYNAHQ	ASTYP	P	V	GDYFASECC

Fig. 2. Alignment of GH3 proteins. Sequence alignment of the regions from GH3 proteins that are involved in the interaction with substrates. A representative member of each of the previously established subgroups is included, plus three proteins from the proposed ninth subgroup. **a** Conserved domains involved in the interaction with the nucleotide. Please note the Phe residues of the GH3.12 proteins in Domain 2. **b** Residues in helix α5 and α6, sheets β8–β9, and residues 230 and 312 (numeration is from CmGH3.12), involved in the interaction with hormone substrates. Accession IDs for these proteins are given in the supplemental data (ESM-1).

		Environmental cues				Energy source	Tissue specificity		Hormones								Transcription factor family				
	Nº	Light / Phytochrome	Circadian rhythms	Biotic stress	Abiotic stress	Carbohydrate Metabolism	Seed / Embryo	Root	ABA	AUX	CK	ET	GA	JA	SA	MYB	AP2	WRKY	bZIP	DoF	Homeo domain
Citrus clementina	49	10	3	4	5	11	4	4	-	8	-	4	2	-	-	6	-	1	-	6	1
Eucalyptus grandis	37	12	8	5	9	3	8	1	3	-	1	1	3	-	1	5	1	1	1	2	1
Juglans regia	58	15	7	3	15	8	8	1	6	1	-	-	1	-	-	4	3	4	3	4	1
Malus domestica	49	17	12	12	8	8	8	7	2	2	-	-	4	-	2	6	2	1	2	2	3
Morus notabilis	44	12	4	4	8	7	8	6	5	2	1	-	1	-	2	7	-	3	3	-	-
Populus euphratica	47	18	10	2	4	6	5	2	2	2	-	5	1	-	-	10	-	1	-	2	-
Populus trichocarpa	57	15	8	11	10	5	7	5	4	3	-	4	3	-	3	12	2	5	2	4	1
Prunus persica	58	7	2	7	11	7	6	5	5	3	-	2	2	2	1	2	2	2	4	4	1
Theobroma cacao	42	10	6	9	1	12	9	3	-	-	-	1	5	-	1	4	-	2	1	1	4
Vitis vinifera	63	15	6	7	11	17	6	3	4	4	-	1	2	1	4	8	2	2	1	2	1

Table 1. Classification of TFBS. Assigned categories for the TFBS (Transcription Factor Binding Sites) identified in the promoter regions of *GH3.12* genes (Subfamily IV) from ten woody species. Total number of identified motifs is given. TFBS are classified according to the environmental stimuli to which they respond, the tissue-specific expression conferred, the hormones influencing their expression, and the identified families of TF putatively binding their promoters. Protein accession IDs are given in ESM-1.